2

THEORETICAL INVESTIGATION OF OPTICAL COMPUTING BASED ON NEURAL NETWORK MODELS

Demetri Psaltis, Xiang-Guang Gu, David Brady
Yaser S. Abu-Mostafa

CALIFORNIA INSTITUTE OF TECHNOLOGY

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
UNLIMITED

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
AFOSR-86-0296

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 88-1185

6a. NAME OF PERFORMING ORGANIZATION
California Institute of Technology

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
AFOSR/NE

6c. ADDRESS *(City, State, and ZIP Code)*
Department of Electrical Engineering
Mail Stop 116-81
Pasadena CA 91125

7b. ADDRESS *(City, State, and ZIP Code)*
Bldg 410
Bolling AFB DC 20332-6448

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
AFOSR/NE

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFOSR-86-0296

8c. ADDRESS *(City, State, and ZIP Code)*
Bldg 410
Bolling AFB DC 20332-6448

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61102F	2305	B1	

11. TITLE *(Include Security Classification)*
Theoretical Investigation of Optical Computing Based on Neural Network Models

12. PERSONAL AUTHOR(S)
Demetri Psaltis

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM 30Sep86 TO 30Sep88

14. DATE OF REPORT *(Year, Month, Day)*
11/17/88

15. PAGE COUNT
18

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The optical implementation of weighted interconnections is investigated and basic relationships are derived between the number of neurons, the number of connections and the size of the optical system that is used to perform the connections. Specific methods for selecting the positions of the neurons to achieve the maximum density of independent connections are presented. The connectivity of a neural network (number of synapses per neuron) is related to the complexity of the problems it can handle. For a network that learns a problem from examples using a local learning rule, it is proved that the entropy of the problem becomes a lower bound for the connectivity of the network.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
□ UNCLASSIFIED/UNLIMITED □ SAME AS RPT. □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL
GILES

22b. TELEPHONE *(Include Area Code)*
(202) 767-4931

22c. OFFICE SYMBOL
NE

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

TABLE OF CONTENTS

2

Final Technical Report

THEORETICAL INVESTIGATION OF OPTICAL COMPUTING BASED ON NEURAL NETWORK MODELS

Demetri Psaltis, Xiang-Guang Gu, David Brady
Yaser S. Abu-Mostafa

Grant AFOSR-86-0296

Submitted to:
Dr. Lee Giles
Air Force Office of Scientific Research
Bolling Air Force Base
Washington, DC

Principal Investigators

Demetri Psaltis and Yaser S. Abu-Mostafa
Department of Electrical Engineering
California Institute of Technology
Pasadena, CA 91125

I

HOLOGRAPHIC IMPLEMENTATIONS OF NEURAL NETWORKS

I.1 INTRODUCTION

One of the attractive features of neural computation is the fact that neural algorithms can be mapped relatively easily onto analog hardware. The use of simple analog devices allows for high parallelism in neural hardware and thus for gains in processing power. Analog VLSI and optics are the two technologies under development for implementations of artificial neural networks. The advantages of VLSI derive from the maturity of silicon device fabrication technology and the sophistication of nonlinear semiconductor devices. The advantages of optical implementations are that three dimensional linear interconnections may be formed and modified relatively easily using optical holography. This is in contrast with the constraints on VLSI which confine integrated networks to two dimensions. For neural network models with a large number of connections per neuron, the area of a VLSI implementation is dominated by the area of the channels which interconnect the processing nodes (the neurons). Optical implementations are typically arranged as shown in Fig.I.1 with planar arrays of neurons interconnected externally to the plane. This architecture permits the area of the plane to be fully populated with active devices, allowing the construction of much larger networks. The disparity between optical and electronic implementations in terms of the number of neurons per unit area which may be realized depends on the density with which the neurons are to be interconnected and the functionality of the neurons.

In this chapter we consider networks with connections which are dense; i.e. each neuron is connected to many others, and irregular; i.e. the strengths of different connections are different. Each neuron is assumed to perform a simple threshold on a weighted sum of the activations of the other neurons to which it is connected. In this case it is not necessary to implement one-to-one connections between any two units. Instead a single "bus" can be used for each neuron that collects the signals from all the units in its receptive field and delivers the accumulated sum to the neuron. This fact reduces dramatically the complexity of the hardware needed to perform the interconnections for both the optical and electronic implementations of neural networks. The simplest VLSI implementation of this architecture is a cross bar which connects M neurons in area M^2 [1]. In this chapter we examine the optical implementation of analog summing interconnections and we derive basic relationships between the number of neurons per unit area at the "neural planes" and the properties of the optical system that is used to perform the connections. In the optical implementations the input port to a "neuron" is a light detector and the output port is an adjacent light source or modulator that is electrically controlled by the the detected signal. The weighted interconnections between the neurons are realized via holograms that are placed between the planes. While our discussion is based on a specific architecture (the Vander Lugt correlator) as an example, the limits we derive and the basic methods we describe are generally applicable with only minor modifications.

I.2 OPTICAL INTERCONNECTIONS USING PLANAR HOLOGRAMS

A schematic diagram of an optical correlator [2] is shown in Fig.I.2. We will utilize this same basic architecture throughout the chapter and consider its implementation with a planar hologram in this section and a volume hologram in the next. A point at the input plane (P_1 in Fig.I.2) is connected to an output point P_2 as follows. The first lens L_1 collimates the light emanating from P_1 into a single plane wave that illuminates the hologram. The direction of propagation of this plane wave has a one-to-one correspondence with the position of P_1 at the input plane. A hologram is placed at the intermediate plane in Fig.I.2. Its purpose is to diffract the incident light towards points at the output plane and thus interconnect input points to output points. We can think of the hologram as a linear superposition of sinusoidal gratings. Each grating diffracts a portion of the incident wave into another plane wave propagating towards the output plane. The difference in the direction of propagation of the incident wave and the direction of the diffracted wave is determined by the spatial frequency and the orientation of the fringes of each grating. The second lens (L_2 in Fig.I.2) converts each of the diffracted waves into a focused spot whose position at the output plane corresponds to the direction of propagation of the diffracted beam. In this manner, each sinusoidal grating that is recorded on the hologram interconnects P_1 to an output point. The weight of the connection is determined by the strength of the recorded grating.

The system of Fig.I.2 is shift invariant. Once the connectivity of a pair of input-output points is determined by recording the appropriate grating on the hologram, then any other input point is connected in the same way to the point at the output plane that is shifted from the original output point by a distance equal to the separation between the two input points. If such a set of four points were selected for the placement of neurons at the input and output planes, then it would not be possible to arbitrarily specify the connectivity between the neurons in the system of Fig.I.2. The strategy that we use in order to provide *independent* interconnections between input and output points is as follows. Once an input/output pair is selected and a grating is recorded for it, then for each additional input location that is used for the placement of a neuron, the point that is shifted by the same amount at the output is excluded from being used as a neuron site. Similarly, a point at the input is eliminated for each additional output point that is used.

This procedure is schematically drawn in Fig.I.3 where the 2-D rectangular grids of available input (top) and output (bottom) pixels are drawn. A grating that connects two points is drawn as a solid arrow in this diagram. The use of a second point automatically connects it to the point at the output marked with an X (dotted arrow) and it is therefore eliminated from the output grid. An analogous diagram is drawn for the use of an additional output point. In general, with reference to Fig.I.3, each grating recorded in the hologram specifies an interconnection between only one pair of input/output points if and only if *the diagram formed by connecting any two input neurons and any two output neurons cannot be a parallelogram.* We now use this criterion to address two issues: a) Capacity, or the maximum number of pixels at the input and output planes that can be used for the placement of neurons and b) The derivation of appropriate sampling grids that provide this maximum capacity.

Let us denote by N_1 (N_2) the number of input (output) neurons and let N be the number of available pixels in 1-D at the input and output planes. The total number of connections that need to be implemented is $N_1 N_2$ and each of these connections must be realized by a distinct grating in order to be independently specifiable. In the diagram of Fig.I.3, each distinct grating corresponds to a vector of a given length and direction. The maximum number of distinct vectors (i.e. each vector having different length and orientation from all others) that can be drawn in the diagram of Fig.I.3, provides us with an upper bound for the number of independent interconnections. We can count how many such vectors there are relatively easily. Pick the point at the lower left corner at the output in Fig.I.3. N^2 distinct vectors can be drawn from this point to points at the input. If we pick any one of the other three output corner points, then each of the N^2 vectors that can be drawn connecting them to points at the input are different except for vectors connecting to points at the perimeter of the input plane. Subtracting these overcounted vectors gives us $4N^2 - 4N + 1$ for the number of distinct vectors. The order of magnitude of the interconnection capacity of this system is therefore

$$N_1 N_2 \leq N^2. \qquad (I.1)$$

For example, let $N_1 = N_2 = 10^4$. Then from Eq.(I.1) we conclude that in order to implement this network we must construct an optical system that is capable of accommodating $N = 10^4$ pixels in 1-D. This applies not only to the input and output planes but also to the hologram, which must have resolution equal to N^2 pixels as well. Notice that the input and output planes are sparsely populated with neurons since only 10^4 out of the available 10^8 pixels are used.

We now describe specific methods for selecting which N_1 (N_2) pixels out of the available N^2 pixels at the input (output) plane to use. This selection can be systematically accomplished in several ways and the resulting sampling grids are not unique. One such pair of sampling grids is shown in Fig.I.4a. In the input a cluster of $N_1 = \sqrt{N} \times \sqrt{N}$ neurons are used whereas at the output the neurons are arranged on a periodic grid with period $\sqrt{N} + 1$. The total number of neurons that can be accommodated at the output is N also. We prove that this is a valid sampling grid by showing that it is impossible to draw a parallelogram on the diagram of Fig.I.4a by connecting any two input points and any two output points. Such a parallelogram cannot be formed because the edge connecting two points on the sampling grid at the input plane would be shorter than $\alpha(\sqrt{N} + 1)$ whereas the edge parallel to it at the output plane would have to be equal to or longer than $\alpha(\sqrt{N} + 1)$. α is a constant that is determined by the orientation of these two edges.

A different sampling grid is shown in Fig.I.4b. The output grid is the same as in the previous case but the input is sampled with period $\sqrt{N}$. We again use the parallelogram test to show that this is a valid sampling grid. The edge of such a parallelogram connecting two input points would have length $\alpha k_1 \sqrt{N}$ with k_1 an integer in the range $0 < k_1 < \sqrt{N}$. The edge of this same parallelogram at the output plane would have length $\alpha k_2(\sqrt{N} + 1)$, with k_2 an integer in the range $0 < k_2 < \sqrt{N}$. The smallest pair of integers that can make the two edges equal is $k_1 = \sqrt{N} + 1$, $k_2 = \sqrt{N}$, which is beyond the range of k_1. Therefore, it is not possible to draw a parallelogram in Fig.I.4b which proves the validity of these sampling grids.

I.3 OPTICAL INTERCONNECTIONS USING VOLUME HOLOGRAMS

We now consider the interconnecting capabilities of the system in Fig.I.2 with a volume rather than a planar hologram in the intermediate plane [3,4,5]. The distinction in the mode of operation between a planar and a volume hologram is the sensitivity of the volume hologram to the angle of incidence of the illumination. We will discuss the angular sensitivity of volume holograms with the help of the k-space diagram of Fig.I.5. The k-space representation is a sphere with radius $2\pi/\lambda$ in which the incident plane wave is drawn as a vector with its origin at the center of the sphere, magnitude equal to $2\pi/\lambda$ and direction that of the incident plane wave. λ is the wavelength of the incident light. The grating is drawn as a vector with its origin the tip of the incident vector, magnitude equal to $2\pi/\Lambda$, and direction pointing perpendicular to the fringes of the grating. Λ is the period of the grating. The diffracted optical wave is drawn as a vector with origin the center of the sphere and magnitude $2\pi/\lambda$. The direction of the diffracted wave is taken to be towards the tip of the grating vector. The efficiency with which light is diffracted is determined by the difference between this diffracted wavevector and the vector formed as the sum of the incident and grating vectors [6]. If the tip of the grating vector falls on the sphere, then this difference reduces to zero and the efficiency is maximized (this is the Bragg condition). For a finite difference the diffraction efficiency is reduced in proportion to the square of the thickness of the crystal, i.e. a thicker crystal is more sensitive to an angular deviation from the Bragg condition.

Returning to Fig.I.2, imagine that a pair of input/output points and a grating have been chosen such that light originating at the input point produces a plane wave that illuminates the hologram at the Bragg angle and the diffracted light is focused at the selected output point. This situation is drawn in the k-space diagram (Fig.I.5) with the diffracted vector being the vectorial sum of the incident and the grating vectors. Consider the two circles that are drawn in Fig.I.5. These circles are formed by the intersection of the k-space sphere with two planes, both of them perpendicular to the grating vector. One of the planes contains the origin and the second contains the tip of the grating vector. Consider an additional incident vector drawn on the k-sphere such that its tip lies on the bottom circle. The grating that is recorded to interconnect the first two neurons is perfectly matched to this additional vector. The direction of the diffracted wave is found by forming the vectorial sum of the additional wavevector and the original grating vector. The tip of the new diffracted wavevector falls on the upper circle. All such incident and diffracted waves define a "degeneracy cone" in k-space along which a single grating specifies the connections of all incident wavevectors that lie on the bottom circle to corresponding diffracted wavevectors on the upper circle. In order to implement independent interconnections in this case, the location of the neurons at the input and output planes must be chosen such that no two input/output pairs share the same degeneracy cone. This condition can be mapped to the input and output planes as shown in Fig.I.6. The grating is drawn as a vector connecting a point at the input to a point at the output and the two circles are approximately mapped to lines perpendicular to the grating vector. In this diagram, the condition that must be obeyed in choosing the location of the neurons is that *the diagram that is formed by connecting any two input neurons and any two output neurons cannot be a rectangle.* As before, we derive the capacity of the correlator implemented with a volume hologram and then present specific algorithms for deriving valid sampling grids.

We can derive an upper bound for the number of independent interconnections that can be implemented with a volume holographic correlator by starting with the connections that the system with the planar hologram can implement and then count the additional connections that are created by the volume hologram. Each distinct vector that can be drawn in Fig.I.6 connecting an input to an output pixel can be used to perform an independent interconnection. From our discussion on planar holograms we know that there are $4N^2 - 4N + 1$ such vectors. With a volume hologram however, each such vector can be used multiple times because when it is translated along the direction of the vector, then it is no longer possible to form a rectangle using the origins and the tips of the original and translated vectors. The maximum number of truly distinct translations we can have for each vector is upper bounded by N, the number of pixels that are available in one dimension. Thus, we obtain the upper bound for the number of independent interconnections that can be implemented with a volume hologram as $(4N^2 - 4N + 1) \times N$, or more simply the order of magnitude is

$$N_1 N_2 \leq N^3. \qquad (I.2)$$

As an example, if $N_1 = N_2 = 10^4$ then using Eq.(I.2) we find $N > 465$. Notice, that this requirement on the space-bandwidth product of the input plane and the optical system is reduced greatly compared to the planar hologram case. As a result it is possible to construct much more compact systems when volume holograms are used. Another way of looking at the distinction between the planar and volume holograms is in terms of the density with which they allow us to populate the input and output planes with neurons. For the symmetric case ($N_1 = N_2$) the number of neurons that can be accommodated by a plane of fixed space-bandwidth product increases by a factor $\sqrt{N}$ when a volume hologram is used.

We now discuss methods for deriving specific sampling grids that achieve the bound of Eq.(I.2). The design criterion that is used in selecting the locations for the placement of neurons at the input and output planes is the avoidance of the formation of a rectangle in the diagram of Fig.I.6, as discussed earlier. The sampling grid shown in Fig.I.7a is constructed by selecting adjacent $\sqrt{N}$ columns, each having N neurons, as the input pattern. The maximum separation in the horizontal direction is $\sqrt{N} - 1$ pixels. If at the output plane any two neurons are separated by less than $\sqrt{N}$ pixels in the horizontal direction, then the possibility exists that a rectangle can be formed at some angle using these two output points and two of the input points. In Fig.I.7a this possibility is eliminated since the output sampling grid consists of $\sqrt{N}$ columns that are separated by $\sqrt{N}$ pixels. Notice that both the input and output planes contain in this case $N^{3/2}$ neurons. A second possibility is shown in Fig.I.7b. In this case the output pattern is the same as the one in Fig.I.7a whereas the input pattern is constructed by $\sqrt{N}$ columns, each being separated from the adjacent column by $\sqrt{N} + 1$ pixels. When we attempt to draw a parallelogram using two input and two output points on the sampling grids of Fig.I.7b, we find that this can only be accomplished if the length of the edge that connects the two input points ($\alpha k_1(\sqrt{N} + 1)$) and the length of the one that connects the two output points ($\alpha k_2 \sqrt{N}$) have equal lengths. k_1 and k_2 are integers and α is a constant. The smallest integers that satisfy this equation are $k_1 = \sqrt{N}$, $k_2 = \sqrt{N} + 1$, which both yield an edge that is larger than N. Hence, it is not possible to form the rectangle within the available $N \times N$ input and output planes and the sampling grids of Fig.I.7b are shown to be valid.

Notice that the number of neurons in either plane for the two sampling grids we presented is $N \times \sqrt{N} = N^{3/2}$. Equivalently, we can think of them as patterns with fractal dimension 3/2. The total number of connections that are implemented by the hologram is $N_1 N_2 = N^3$. Comparing this result with Eq.(I.2) we find that these sampling grids provide the full interconnection capacity that is available with a volume hologram.

I.4 CONCLUSION

We described how holograms can be used to provide arbitrary, full interconnection between two planes of neurons. The methods presented can be extended in relatively straightforward ways to design other sampling grids and to realize non-symmetric (i.e. $N_1 \neq N_2$) and local interconnections [7]. All such sampling grids share the property that the available degrees of freedom of the hologram are fully utilized. In the case of planar holograms there are N^2 pixels available in the area of the hologram where as 3-D storage in volume holograms increases the capacity to N^3. The overall volume required is in both cases proportional to N^3 (obtained as the product of the area of each plane which is proportional to N^2 and the minimum separation between planes which is proportional to N). In Table 1 we compare a planar versus a thick hologram in terms of the size of the optical system required to fully interconnect two layers each having $N_1 = N_2 = M$ neurons. The required overall volume of the system is M times smaller when a volume hologram is used. It should be pointed out however that the reduction in system volume that results from the 3-D storage capability of volume holograms is accompanied by a reduction in the degree of control we have in storing information. This is due to the fact that while information is stored throughout the three dimensional medium in a volume hologram, we can only affect the stored contents through information that we specify on the two dimensional surface that encloses the hologram [8]. The consequences of this fact [9] must be included along with the geometrical arguments presented here for a complete assessment of the relative merits of the two types of holographic interconnections.

Table 1

PLANAR VS. VOLUME HOLOGRAMS		
M = Number of Neurons		
N = 1-D Space Bandwidth Product		
	2-D	3-D
Linear Dimension	M	$M^{2/3}$
Area	M^2	$M^{4/3}$
Total System Volume	M^3	M^2
Volume Ratio	$R = V_{2-D}/V_{3-D} = M$	

REFERENCES

[1] L. D. Jackel, R. E. Howard, H. P. Graf, B. Straughn, and J. S. Denker, "Artificial neural networks for computing", J. Vac. Sci. Technol., **B 4**, 61(1986).

[2] A. B. Vander Lugt, IEEE Trans. Inform. Theory, **IT-10**(2), 98(1985).

[3] P. J. van Heerden, "Theory of optical information storage in solids", Appl. Opt., **2**, 393(1963).

[4] H. Lee, X. Gu, and D. Psaltis, "Volume holographic interconnections with maximal capacity and minimal crosstalk", to appear in J. Appl. Phys..

[5] D. Psaltis, J. Yu, X. Gu, and H. Lee, "Optical neural nets implemented with volume holograms", *Technical Digest of Topical Meeting on Optical Computing*, Optical Society of America, Washington, DC, (1987).

[6] H. Kogelnik, "Coupled Wave Theory for Thick Hologram Gratings", Bell Sys. Tech. Journal, **48**(9),2909(1969).

[7] X. Gu and D. Psaltis, "Local and asymmetric interconnections using volume holograms", OSA Annual Meeting, *1988 Technical Digest Series*, **11**, Optical Society of America, Washington, DC, 148(1988).

[8] S. Hudson, D. J. Brady, and D. Psaltis, "Properties of 3-D imaging systems", OSA Annual Meeting, *1988 Technical Digest Series*, **11**, Optical Society of America, Washington, DC, 74(1988).

[9] D. Psaltis, D. J. Brady and K. Wagner, "Adaptive optical networks using photorefractive crystals", Appl. Opt., **27**(9), 1752(1988).

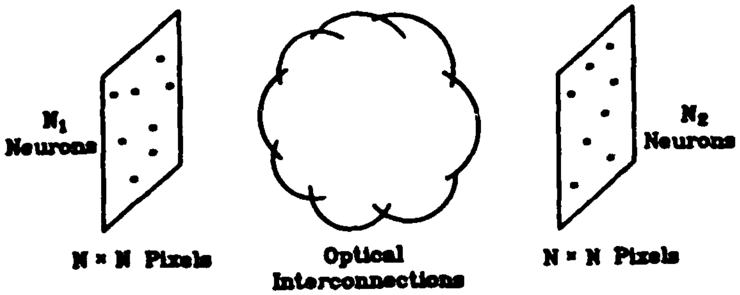

Fig.I.1 Basic optical processing system.

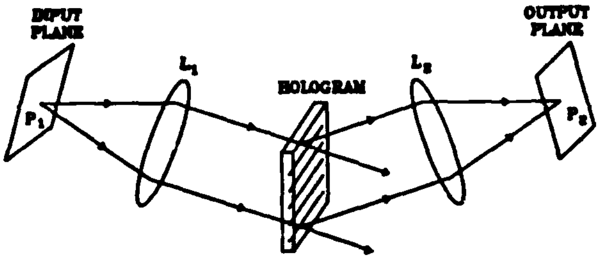

Fig.I.2 Vander Lugt correlator.

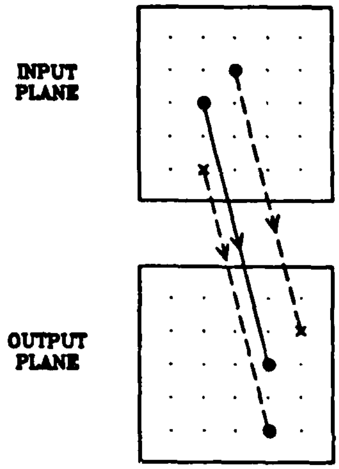

Fig.I.3 For 2-D holograms, no figure with vertices at a pair of input neurons and a pair of output neurons may form a parallelogram.

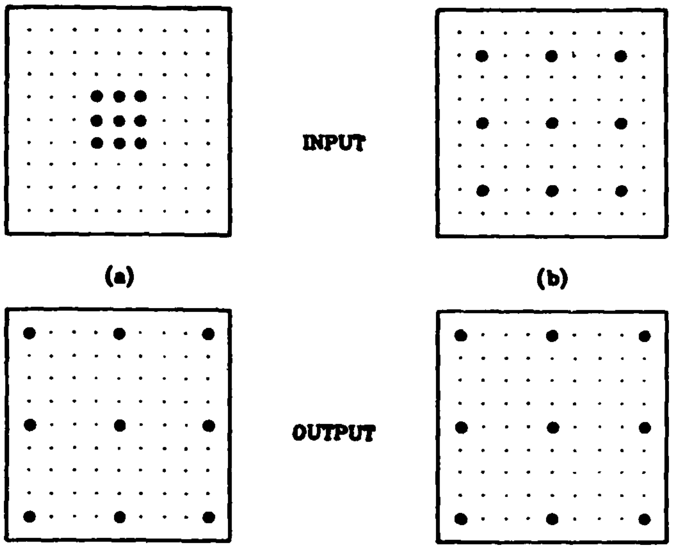

Fig.I.4 Sampling grids for planar holograms.

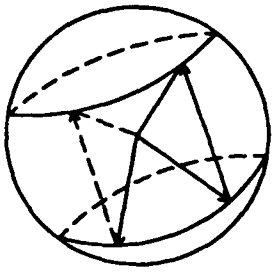

Fig.I.5 k-space diagram.

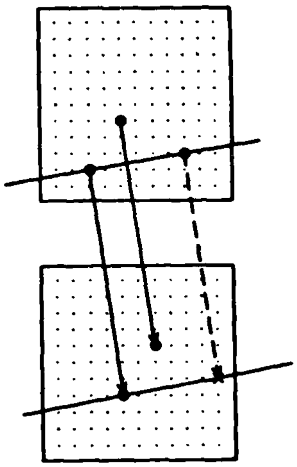

Fig.I.6 For 3-D holograms, no figure with vertices at a pair of input neurons and a pair of output neurons may form a rectangle.

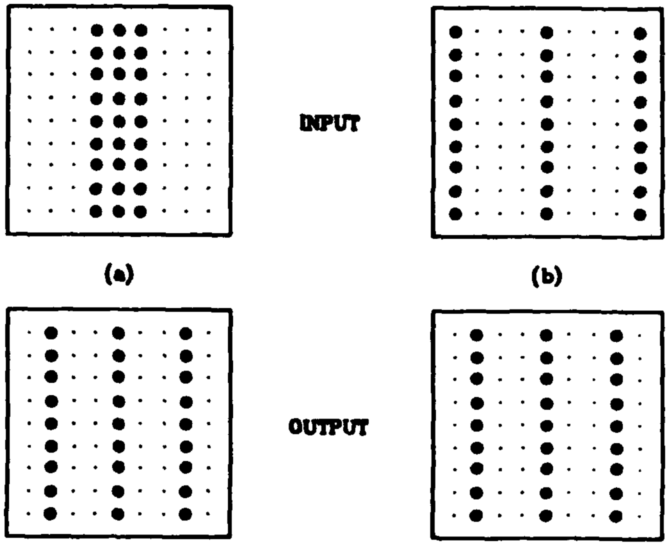

Fig.I.7 Sampling grids for volume holograms.

II

LOWER BOUND FOR CONNECTIVITY IN LOCAL-LEARNING NEURAL NETWORKS

II.1 INTRODUCTION

Learning by example has emerged as the most important question in neural networks. Clearly, a given neural network cannot just learn any function, there must be some restrictions on which networks can learn which functions. One obvious restriction, which is independent of the learning aspect, is that the network must be big enough to accommodate the circuit complexity of the function it will eventually simulate. A restriction that arises merely from the fact that the network is expected to *learn* the function, rather than being purposely designed for the function is reported in [Abu-Mostafa, 1988]. The restriction imposes a lower bound on the connectivity of the network (number of synapses per neuron). In this paper, we describe a generalization of this result by removing one of the requirements on the learning mechanism. Instead of requiring that the training sample itself be loaded directly into the neurons, we now allow arbitrary features to be extracted from the sample and loaded into the neurons. This also implies that the number of neurons can be very large with respect to the number of bits in each sample.

However, our generalized result still assumes a local-learning mechanism. The local-learning assumption allows only local information to be used by each neuron in its learning effort. The assumption cannot be completely removed since a powerful learning mechanism can be designed that will find one of the low-connectivity (e.g., two-input-NAND-gate) circuits that fits all the training samples, perhaps by exhaustive search. Local-learning is a strong assumption that excludes sophisticated learning mechanisms used in neural-network models.

The lower bound on the connectivity of the network is given in terms of the *entropy* of the environment that provides the training samples. Entropy is a quantitative measure of the disorder or randomness in an environment or, equivalently, the amount of information needed to specify the environment. In section 2, we shall introduce the formal definitions and results, but we start here with an informal exposition of the ideas involved.

The environment in our model produces patterns represented by N bits $\mathbf{x} = x_1 \cdots x_N$ (pixels in the picture of a visual scene if you will). Only h different patterns can be generated by a given environment, where $h < 2^N$ (the entropy is essentially $\log_2 h$). No knowledge is assumed about which patterns the environment can generate, only that there are h of them. In the learning process, a number of sample patterns are generated at random from the environment. A large number of binary features are extracted from each sample and input to the network, one feature per neuron. The network uses this information to set its internal parameters and gradually tune itself to this particular environment. Because of the network architecture, each neuron knows only its own bit and the bits of the neurons it is directly connected to by a synapse. Hence, the learning rules are local: a neuron does not have the benefit of the entire global pattern that is being learned.

After the learning process has taken place, each neuron is ready to perform a function *defined by what it has learned.* The collective interaction of the functions of the neurons is what defines the overall function of the network. The main result of this paper is that (roughly speaking) if the connectivity of the network is less than the entropy of the environment, the network cannot learn about the environment. The idea of the proof is to show that if the connectivity is small, the final function of each neuron is independent of the environment, and hence to conclude that the overall network has accumulated no information about the environment it is supposed to learn about.

II.2 LOCAL-LEARNING NETWORKS

A neural network can be described as an undirected graph (the vertices are the neurons and the edges are the synapses). Label the neurons $1,\cdots,\mathcal{N}$. Each neuron can store one bit at a time, but it also has access to those bits stored by the other neurons to which it is directly connected by a synapse. By local learning, we mean that the adjustments a neuron makes when the network is loaded with a training sample will depend only on the bits it has access to, namely its own bit and the bits of its neighbors. In other words, the neuron does not have the benefit of the global picture in its effort to learn, just the bits it can see locally.

During the learning phase, an unknown environment provides a sequence of training samples to the network. The environment is a subset $e \subseteq \{0,1\}^N$ (each $\mathbf{x} \in e$ is a possible sample from the environment). When the environment produces a sample $\mathbf{x}$, binary features $f_1,\cdots,f_{\mathcal{N}}$ are extracted from $\mathbf{x}$ and loaded into the neurons $1,\cdots,\mathcal{N}$, respectively (a feature is a function $f_i : \{0,1\}^N \rightarrow \{0,1\}$). For a given network, the features $f_1,\cdots,f_{\mathcal{N}}$ are arbitrary but fixed, and $\mathcal{N}$ (the number of neurons) can be much larger than N (the number of bits in a sample), e.g., $\mathcal{N}$ can be superexponential in N.

As the samples from the unknown environment e come in, each neuron sees the subset of features carried by itself and its neighbors. Consider an arbitrary neuron that sees K features (we will assume $K \leq N \leq \mathcal{N}$ throughout), and relabel $1,\cdots,\mathcal{N}$ to make these features $f_1,\cdots,f_K$. Based on the values $f_1,\cdots,f_K$ assume as $\mathbf{x}$ varies over e, the neuron is supposed to learn about the environment such that, after the learning phase is over, the collective behaviour of the network is tuned to the environment e that provided the samples. How the neurons absorb the learning information and what computation the network is supposed to perform eventually are left deliberately unspecified. The arguments in this paper are based on the lack of information rather than the failure to use information.

The connectivity is measured by the parameter K. Since our result is asymptotic in N, we will specify K as a function of N; $K = \alpha N$ where $\alpha = \alpha(N)$ satifies $\lim_{N\rightarrow\infty} \alpha(N) = \alpha_o$ $(0 < \alpha_o < 1)$. To formalize the concept of unknown environment, we will consider the ensemble of environments $\mathcal{E}$ of fixed entropy [Abu-Mostafa, 1986]

$$\mathcal{E} = \mathcal{E}(N) = \left\{ e \subseteq \{0,1\}^N \;\mid\; |e| = h \right\}$$

where $h = 2^{\beta N}$ (the entropy is essentially $\log_2 h = \beta N$) and $\beta = \beta(N)$ satisfies $\lim_{N\rightarrow\infty} \beta(N) = \beta_o$ $(0 < \beta_o < 1)$. The probability distribution on $\mathcal{E}$ is uniform; any environment $e \in \mathcal{E}$ is as likely to occur as any other.

The neuron sees only the K (fixed but arbitrary) functions $f_1, \cdots, f_K$ of each $\mathbf{x}$ generated by the environment e. For each e, we define the function $n : \{0,1\}^K \to \{0,1,2,\cdots\}$ where

$$n(a_1 \cdots a_K) = |\{\mathbf{x} \in e \ \mid \ f_k(\mathbf{x}) = a_k \text{ for } k = 1, \cdots, K\}|$$

and the normalized version

$$\nu(a_1 \cdots a_K) = \frac{n(a_1 \cdots a_K)}{h}$$

The function ν describes the relative frequency of occurrence for each of the 2^K binary vectors $f_1(\mathbf{x}) \cdots f_K(\mathbf{x})$ as $\mathbf{x}$ runs through all h vectors in e. In other words, ν specifies the nonlinear projection of e as seen by the neuron. Clearly, $\nu(\mathbf{a}) \geq 0$ for all $\mathbf{a} \in \{0,1\}^K$ and $\sum_{\mathbf{a} \in \{0,1\}^K} \nu(\mathbf{a}) = 1$.

Corresponding to two environments e_1 and e_2, we will have two functions ν_1 and ν_2. If ν_1 is not distinguishable from ν_2, the neuron cannot tell the difference between e_1 and e_2. The distinguishability between ν_1 and ν_2 can be measured by

$$d(\nu_1, \nu_2) = \frac{1}{2} \sum_{\mathbf{a} \in \{0,1\}^K} |\nu_1(\mathbf{a}) - \nu_2(\mathbf{a})|$$

The range of $d(\nu_1, \nu_2)$ is $0 \leq d(\nu_1, \nu_2) \leq 1$, where '0' corresponds to complete indistinguishability while '1' corresponds to maximum distinguishability. The main result of this paper is to relate this distinguishability to how the connectivity of the network compares with the entropy of the environment.

II.3 MAIN RESULT

Let e_1 and e_2 be independently selected environments from $\mathcal{E}$ according to the uniform probability distribution. $d(\nu_1, \nu_2)$ is now a random variable, and we are interested in the expected value $E(d(\nu_1, \nu_2))$. The case where $E(d(\nu_1, \nu_2)) = 0$ corresponds to the neuron getting no information about the environment, while the case where $E(d(\nu_1, \nu_2)) = 1$ corresponds to the neuron getting maximum information. $E(d(\nu_1, \nu_2))$ depends, among other things, on the choice of the features $f_1, \cdots, f_K$. For example, a poor choice of the f_k's as constant functions forces $E(d(\nu_1, \nu_2))$ to be zero regardless of K. For which values of K does there exist a choice of the f_k's that makes $E(d(\nu_1, \nu_2))$ close to 1, and for which values is $E(d(\nu_1, \nu_2))$ close to 0 for all choices of the f_k's? The theorem predicts these extremes depending on how the connectivity (represented by α_o in the limit) compares with the entropy (represented by β_o in the limit).

Theorem.

1. If $\alpha_o > \beta_o$, then for every N there exist functions $f_1, \cdots, f_K$, such that $\lim_{N \to \infty} E\left(d(\nu_1, \nu_2)\right) = 1$.
2. If $\alpha_o < \beta_o$, then for all functions $f_1, \cdots, f_K$ for all N, $\lim_{N \to \infty} E\left(d(\nu_1, \nu_2)\right) = 0$.

Proof.

1. We shall take the functions $f_1, \cdots, f_K$ to be the simple projection functions $f_k(x_1 \cdots x_k \cdots x_N) = x_k$. Thus the neuron sees the first K bits $x_1 \cdots x_K$ of the sample $\mathbf{x} = x_1 \cdots x_N$. We start with some basic properties about the ensemble of environments $\mathcal{E}$. Since the probability distribution on $\mathcal{E}$ is uniform and since $|\mathcal{E}| = \binom{2^N}{h}$, we have

$$\Pr(e) = \binom{2^N}{h}^{-1}$$

which is equivalent to generating e by choosing h elements $\mathbf{x} \in \{0,1\}^N$ with uniform probability (without replacement). It follows that

$$\Pr(\mathbf{x} \in e) = \frac{h}{2^N}$$

while for $\mathbf{x}_1 \neq \mathbf{x}_2$,

$$\Pr(\mathbf{x}_1 \in e \ , \ \mathbf{x}_2 \in e) = \frac{h}{2^N} \times \frac{h-1}{2^N - 1}$$

and so on.

The functions n and ν are defined on K-bit vectors. For the above choice of the functions $f_1 \cdots f_K$, the statistics of $n(\mathbf{a})$ (a random variable for fixed $\mathbf{a}$) is independent of $\mathbf{a}$

$$Pr(n(\mathbf{a}_1) = m) = Pr(n(\mathbf{a}_2) = m)$$

which follows from the symmetry with respect to each bit of $\mathbf{a}$. The same holds for the statistics of $\nu(\mathbf{a})$. The expected value $E(n(\mathbf{a})) = h2^{-K}$ (h objects going into 2^K cells), hence $E(\nu(\mathbf{a})) = 2^{-K}$.

We expand $E\left(d(\nu_1, \nu_2)\right)$ as follows

$$\begin{aligned} E\left(d(\nu_1,\nu_2)\right) &= E\left(\frac{1}{2} \sum_{\mathbf{a}\in\{0,1\}^K} |\nu_1(\mathbf{a}) - \nu_2(\mathbf{a})|\right) \\ &= \frac{1}{2h} \sum_{\mathbf{a}\in\{0,1\}^K} E\left(|n_1(\mathbf{a}) - n_2(\mathbf{a})|\right) \\ &= \frac{2^K}{2h} E(|n_1 - n_2|) \end{aligned}$$

where n_1 and n_2 denote $n_1(0 \cdots 0)$ and $n_2(0 \cdots 0)$, respectively, and the last step follows from the fact that the statistics of $n_1(\mathbf{a})$ and $n_2(\mathbf{a})$ is independent of $\mathbf{a}$. Therefore, to prove the first part of the theorem, we assume $\alpha_o > \beta_o$ and evaluate $E(|n_1 - n_2|)$ for large N. Let n denote $n(0 \cdots 0)$, and consider $\Pr(n = 0)$. For n to be zero, all 2^{N-K} strings $\mathbf{x}$ of N bits starting with K 0's must *not* be in the environment e. Hence

$$\Pr(n = 0) = (1 - \frac{h}{2^N})(1 - \frac{h}{2^N - 1}) \cdots (1 - \frac{h}{2^N - 2^{N-K} + 1})$$

where the first term is the probability that $0\cdots00 \notin e$, the second term is the probability that $0\cdots01 \notin e$ given that $0\cdots00 \notin e$, and so on.

$$\begin{aligned}
&\geq \left(1 - \frac{h}{2^N - 2^{N-K}}\right)^{2^{N-K}} \\
&= \left(1 - h2^{-N}(1-2^{-K})^{-1}\right)^{2^{N-K}} \\
&\geq (1 - 2h2^{-N})^{2^{N-K}} \\
&\geq 1 - 2h2^{-N}2^{N-K} \\
&= 1 - 2h2^{-K}
\end{aligned}$$

Hence, $\Pr(n_1 = 0) = \Pr(n_2 = 0) = \Pr(n = 0) \geq 1 - 2h2^{-K}$. However, $E(n_1) = E(n_2) = h2^{-K}$. Therefore,

$$\begin{aligned}
E(|n_1 - n_2|) &= \sum_{i=0}^{h}\sum_{j=0}^{h} \Pr(n_1 = i, n_2 = j)|i - j| \\
&= \sum_{i=0}^{h}\sum_{j=0}^{h} \Pr(n_1 = i)\Pr(n_2 = j)|i - j| \\
&\geq \sum_{j=0}^{h} \Pr(n_1 = 0)\Pr(n_2 = j)j \\
&\quad + \sum_{i=0}^{h} \Pr(n_1 = i)\Pr(n_2 = 0)i
\end{aligned}$$

which follows by throwing away all the terms where neither i nor j is zero (the term where both i an j are zero appears twice for convenience, but this term is zero anyway).

$$\begin{aligned}
&= \Pr(n_1 = 0)E(n_2) + \Pr(n_2 = 0)E(n_1) \\
&\geq 2(1 - 2h2^{-K})h2^{-K}
\end{aligned}$$

Substituting this estimate in the expression for $E(d(\nu_1, \nu_2))$, we get

$$\begin{aligned}
E(d(\nu_1, \nu_2)) &= \frac{2^K}{2h}E(|n_1 - n_2|) \\
&\geq \frac{2^K}{2h} \times 2(1 - 2h2^{-K})h2^{-K} \\
&= 1 - 2h2^{-K} \\
&= 1 - 2 \times 2^{(\beta - \alpha)N}
\end{aligned}$$

Since $\alpha_o > \beta_o$ by assumption, this lower bound goes to 1 as N goes to infinity. Since 1 is also an upper bound for $d(\nu_1, \nu_2)$ (and hence an upper bound for the expected value $E(d(\nu_1, \nu_2))$), $\lim_{N\to\infty} E(d(\nu_1, \nu_2))$ must be 1.

2. Assume $\alpha_o < \beta_o$, and consider arbitrary functions $f_1, \cdots, f_K$. Define

$$\bar{n}(\mathbf{a}) = \frac{h}{2^N} \left| \{ \mathbf{x} \in \{0,1\}^N \mid f_k(\mathbf{x}) = a_k \text{ for } k = 1, \cdots, K \} \right|$$

We expand $E(d(\nu_1, \nu_2))$ as follows

$$\begin{aligned}
E(d(\nu_1,\nu_2)) &= \frac{1}{2h} \sum_{\mathbf{a}\in\{0,1\}^K} E(|n_1(\mathbf{a}) - n_2(\mathbf{a})|) \\
&= \frac{1}{2h} \sum_{\mathbf{a}\in\{0,1\}^K} E\left(|(n_1(\mathbf{a}) - \bar{n}(\mathbf{a})) - (n_2(\mathbf{a}) - \bar{n}(\mathbf{a}))|\right) \\
&\leq \frac{1}{2h} \sum_{\mathbf{a}\in\{0,1\}^K} E(|n_1(\mathbf{a}) - \bar{n}(\mathbf{a})| + |n_2(\mathbf{a}) - \bar{n}(\mathbf{a})|) \\
&= \frac{1}{2h} \sum_{\mathbf{a}\in\{0,1\}^K} E(|n_1(\mathbf{a}) - \bar{n}(\mathbf{a})|) + E(|n_2(\mathbf{a}) - \bar{n}(\mathbf{a})|) \\
&= \frac{1}{h} \sum_{\mathbf{a}\in\{0,1\}^K} E(|n(\mathbf{a}) - \bar{n}(\mathbf{a})|)
\end{aligned}$$

The statistics of $n(\mathbf{a})$ now depends on $\mathbf{a}$ since the functions $f_1 \cdots f_K$ are arbitrary. To evaluate $E(|n(\mathbf{a}) - \bar{n}(\mathbf{a})|)$, we first show that $\bar{n}(\mathbf{a}) = E(n(\mathbf{a}))$, then estimate the variance of $n(\mathbf{a})$ and use the fact that $E(|n(\mathbf{a}) - E(n(\mathbf{a}))|) \leq \sqrt{\text{var}(n(\mathbf{a}))}$. We write

$$n(\mathbf{a}) = \sum_{\mathbf{x}\in\{0,1\}^N} \delta(\mathbf{x},\mathbf{a})\delta(\mathbf{x})$$

where $\delta(\mathbf{x},\mathbf{a}) = 1$ if $f_k(\mathbf{x}) = a_k$ for $k = 1, \cdots, K$ and is zero otherwise, and $\delta(\mathbf{x}) = 1$ if $\mathbf{x} \in e$ and is zero otherwise (while $\delta(\mathbf{x},\mathbf{a})$ is fixed for given $\mathbf{x}$ and $\mathbf{a}$, $\delta(\mathbf{x})$ is a random variable for a given $\mathbf{x}$). Hence

$$E(n(\mathbf{a})) = \sum_{\mathbf{x}\in\{0,1\}^N} \delta(\mathbf{x},\mathbf{a})E(\delta(\mathbf{x}))$$

The expected value of $\delta(\mathbf{x})$ is $\Pr(\mathbf{x} \in e) = h/2^N$. Factoring this out, we are left with $\sum_{\mathbf{x}\in\{0,1\}^N} \delta(\mathbf{x},\mathbf{a})$ which equals $\left|\{\mathbf{x} \in \{0,1\}^N \mid f_k(\mathbf{x}) = a_k \text{ for } k = 1, \cdots, K\}\right|$, hence $E(n(\mathbf{a}))$ indeed equals $\bar{n}(\mathbf{a})$.

Since $\text{var}(n(\mathbf{a})) = E((n(\mathbf{a}))^2) - (E(n(\mathbf{a})))^2$, we need an estimate for $E((n(\mathbf{a}))^2)$.

$$\begin{aligned}
E((n(\mathbf{a}))^2) &= E\left(\sum_{\mathbf{x}_1\in\{0,1\}^N} \sum_{\mathbf{x}_2\in\{0,1\}^N} \delta(\mathbf{x}_1,\mathbf{a})\delta(\mathbf{x}_2,\mathbf{a})\delta(\mathbf{x}_1)\delta(\mathbf{x}_2) \right) \\
&= \sum_{\mathbf{x}_1}\sum_{\mathbf{x}_2} \delta(\mathbf{x}_1,\mathbf{a})\delta(\mathbf{x}_2,\mathbf{a})E(\delta(\mathbf{x}_1)\delta(\mathbf{x}_2))
\end{aligned}$$

For the 'diagonal' terms ($\mathbf{x}_1 = \mathbf{x}_2$), we get $\sum_{\mathbf{x}} \delta(\mathbf{x}, \mathbf{a}) E(\delta(\mathbf{x}))$ (since $\delta^2 = \delta$), which equals $\bar{n}(\mathbf{a})$. For the 'off-diagonal' terms ($\mathbf{x}_1 \neq \mathbf{x}_2$), we get

$$\begin{aligned}
&\sum_{\mathbf{x}_1} \sum_{\mathbf{x}_2 \neq \mathbf{x}_1} \delta(\mathbf{x}_1, \mathbf{a}) \delta(\mathbf{x}_2, \mathbf{a}) E(\delta(\mathbf{x}_1)\delta(\mathbf{x}_2)) \\
&= \sum_{\mathbf{x}_1} \sum_{\mathbf{x}_2 \neq \mathbf{x}_1} \delta(\mathbf{x}_1, \mathbf{a}) \delta(\mathbf{x}_2, \mathbf{a}) \Pr(\mathbf{x}_1 \in e, \mathbf{x}_2 \in e) \\
&= \sum_{\mathbf{x}_1} \sum_{\mathbf{x}_2 \neq \mathbf{x}_1} \delta(\mathbf{x}_1, \mathbf{a}) \delta(\mathbf{x}_2, \mathbf{a}) \frac{h(h-1)}{2^N(2^N-1)} \\
&= \sum_{\mathbf{x}_1} \sum_{\mathbf{x}_2} \delta(\mathbf{x}_1, \mathbf{a}) \delta(\mathbf{x}_2, \mathbf{a}) \frac{h(h-1)}{2^N(2^N-1)} - \sum_{\mathbf{x}} \delta(\mathbf{x}, \mathbf{a}) \frac{h(h-1)}{2^N(2^N-1)}
\end{aligned}$$

The last step follows by adding and subtracting the missing terms of the double summation. Noting that $\bar{n}(\mathbf{a}) = \frac{h}{2^N} \sum_{\mathbf{x}} \delta(\mathbf{x}, \mathbf{a})$, this can be rewritten as

$$\frac{2^N(h-1)}{h(2^N-1)} (\bar{n}(\mathbf{a}))^2 - \frac{h-1}{2^N-1} \bar{n}(\mathbf{a})$$

Putting the contributions from the diagonal and off-diagonal terms together, we get

$$\begin{aligned}
E((n(\mathbf{a}))^2) &= \bar{n}(\mathbf{a}) + \frac{2^N(h-1)}{h(2^N-1)} (\bar{n}(\mathbf{a}))^2 - \frac{h-1}{2^N-1} \bar{n}(\mathbf{a}) \\
&= \frac{2^N(h-1)}{h(2^N-1)} (\bar{n}(\mathbf{a}))^2 + \frac{2^N-h}{2^N-1} \bar{n}(\mathbf{a}) \\
\mathrm{var}(n(\mathbf{a})) &= E((n(\mathbf{a}))^2) - (E(n(\mathbf{a})))^2 \\
&= \frac{2^N(h-1)}{h(2^N-1)} (\bar{n}(\mathbf{a}))^2 + \frac{2^N-h}{2^N-1} \bar{n}(\mathbf{a}) - (\bar{n}(\mathbf{a}))^2 \\
&= \frac{h-2^N}{h(2^N-1)} (\bar{n}(\mathbf{a}))^2 + \frac{2^N-h}{2^N-1} \bar{n}(\mathbf{a}) \\
&= \frac{2^N-h}{2^N-1} \bar{n}(\mathbf{a}) \left(1 - \frac{1}{h} \bar{n}(\mathbf{a})\right) \\
&\leq \frac{2^N-h}{2^N-1} \bar{n}(\mathbf{a}) \\
&\leq \bar{n}(\mathbf{a})
\end{aligned}$$

Thus we have $E(|n(\mathbf{a}) - \bar{n}(\mathbf{a})|) \leq \sqrt{\mathrm{var}(n(\mathbf{a}))} \leq \sqrt{\bar{n}(\mathbf{a})}$. Now, we rewrite the estimate for $E(d(\nu_1, \nu_2))$

$$\begin{aligned}
E(d(\nu_1, \nu_2)) &\leq \frac{1}{h} \sum_{\mathbf{a} \in \{0,1\}^K} E(|n(\mathbf{a}) - \bar{n}(\mathbf{a})|) \\
&\leq \frac{1}{h} \sum_{\mathbf{a} \in \{0,1\}^K} \sqrt{\bar{n}(\mathbf{a})}
\end{aligned}$$

The values of the individual $\bar{n}(\mathbf{a})$ will depend on the choice of $f_1 \cdots f_K$. However, $\sum_{\mathbf{a}\in\{0,1\}^K} \bar{n}(\mathbf{a})$ always equals h (from the definition of $\bar{n}(\mathbf{a})$). Therefore, one can obtain an upper bound for $E(d(\nu_1, \nu_2))$ by maximizing $\sum_{\mathbf{a}\in\{0,1\}^K} \sqrt{\bar{n}(\mathbf{a})}$ subject to $\sum_{\mathbf{a}\in\{0,1\}^K} \bar{n}(\mathbf{a}) = h$. The maximum occurs when all $\bar{n}(\mathbf{a})$ are equal $(= h2^{-K})$. Hence, $E(d(\nu_1, \nu_2)) \leq \frac{1}{h} 2^K \sqrt{h2^{-K}} = \sqrt{\frac{2^K}{h}} = 2^{\frac{1}{2}(\alpha-\beta)N}$. Since $\alpha_o < \beta_o$ by assumption, this upper bound goes to 0 as N goes to infinity. Since 0 is also a lower bound for $d(\nu_1, \nu_2)$ (and hence a lower bound for the expected value $E(d(\nu_1, \nu_2))$), $\lim_{N\to\infty} E(d(\nu_1, \nu_2))$ must be 0. ■

II.4 CONCLUSION

We have shown that, under the assumption of local learning, each neuron must have at least a certain number of synapses in order to be able to distinguish between environments based on the statistics of information it sees. While the result is expressed as a limit, it is seen in the proof that the rate of convergence to this limit is exponential in N, the dimensionality of the problem. Further work should address the weakening of the local-learning assumption, perhaps by restricting the amount of global information flow or by restricting the ability of the neuron to make use of the information it sees (e.g., by modeling its learning mechanism as a finite-state machine).

REFERENCES

Abu-Mostafa, Y. S. (1986), The complexity of information extraction, *IEEE Trans. on Information Theory* **IT-32**, 513-525.

Abu-Mostafa, Y. S. (1988), Connectivity versus Entropy, *in* "Neural Information Processing Systems," D. Anderson (ed.), *American Institute of Physics*.